Indonesian Journal of Electrical Engineering and Computer Science
Vol. 17, No. 3, March 2020, pp. 1210~1214
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v17.i3.pp1210-1214

A malicious URLs detection system using optimization and machine learning classifiers

Ong Vienna Lee[1], Ahmad Heryanto[2], Mohd Faizal Ab Razak[3], Anis Farihan Mat Raffei[4], Danakorn Nincarean Eh Phon[5], Shahreen Kasim[6], Tole Sutikno[7]
[1,3,4,5]Faculty of Computer Systems & Software Engineering, University Malaysia Pahang, Malaysia
[2]Department of Computer Engineering, Universitas Sriwijaya, Indonesia
[6]Faculty of Computer Science & Information Technology, Universiti Tun Hussein Onn Malaysia, Malaysia
[7]Department of Electrical and Computer Engineering, Universitas Ahmad Dahlan, Indonesia

Article Info

Article history:

Received Jul 29, 2019
Revised Sep 20, 2019
Accepted Oct 11, 2019

Keywords:

Android
Detection system
Features optimization
Machine learning
URLs

ABSTRACT

The openness of the World Wide Web (Web) has become more exposed to cyber-attacks. An attacker performs the cyber-attacks on Web using malware Uniform Resource Locators (URLs) since it widely used by internet users. Therefore, a significant approach is required to detect malicious URLs and identify their nature attack. This study aims to assess the efficiency of the machine learning approach to detect and identify malicious URLs. In this study, we applied features optimization approaches by using a bio-inspired algorithm for selecting significant URL features which able to detect malicious URLs applications. By using machine learning approach with static analysis technique is used for detecting malicious URLs applications. Based on this combination as well as significant features, this paper shows promising results with higher detection accuracy. The bio-inspired algorithm: particle swarm optimization (PSO) is used to optimized URLs features. In detecting malicious URLs, it shows that naïve Bayes and support vector machine (SVM) are able to achieve high detection accuracy with rate value of 99%, using URL as a feature.

Copyright © 2020 Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

Ahmad Firdaus,
Faculty of Computer Systems & Software Engineering,
University Malaysia Pahang,
Lebuhraya Tun Razak, 26300 Gambang, Kuantan, Pahang, Malaysia.
Email: firdausza@ump.edu.my

1. INTRODUCTION

A webpages services increasingly prevail; causing business and peoples move toward web applications. At present, most people highly depend on web applications for routine activities such as communications, internet banking, online shopping, information gathering, forum discussion and socializing. The increasing of web applications exposed to the various threat that exploit their vulnerabilities [1, 2]. An attacker used web application vulnerabilities as a stepping stone to compromised URLs for hideous purposes [3, 4]. For instance, attackers used URL to perform an attack on websites. Attackers insert a redirect code into a compromised URLs so that the user will be navigated automatically to malicious URLs [5-7]. This malicious URLs also redirect the user to download a malicious application such as botnet into a computer and cause attacker able to collect confidential information such as banking number and contact information [8, 9].

Malicious URLs continuous to grow and there are 230,000 new malware samples per day [5]. According to Cybint News, the attackers launch their attack for every 39 seconds and have infected 64% of companies [10]. Due to this attack, Kaspersky Lab Solution has blocked more than 7 million attacks and recognizes 282,807,433 unique URLs as a malicious [11]. These malicious attempts to collect confidential

information such as account number and password to steal money from computers users. In addition, in February 2018 there is biggest attack launched by attackers, where 100 of WordPress and Joomla sites infected by malware known as ionCube [12]. These figures show that attackers used almost any vulnerability within URLs applications in order to perform an attack and exploit based attack [13]. To solve this problem, the users try to shield the computer by updating the version of websites applications, the anti-malware and in the midst of doing so, the computer user has to give constant attention to the accessed URLs application.

Recently, studies have shown that there is a number of detection approaches available to combat the increasing number of malicious URLs. For an instant, the signature-based and behavioral based technique [14] is used to detect malicious URLs [15, 16]. In particular, the signature analysis aims at detecting malicious URLs by analyzing their signature. This signature stored in the database repository which represents all of the knowledge the signature-based approach has, as it concerns to malicious URLs detection. Furthermore, the behavioral analysis, also known as heuristic analysis detects malicious URLs by investigating the program in an isolated environment. Other than that, the heuristic analysis applies a machine learning approach and data mining to learn the behavior of executable malicious URLs. Besides, an identification of the most appropriate set of features: URL, host, content, graph and blacklist, help in efficiently distinguishing web pages and URLs into malicious benign.

Although many security defenses are developed against malicious URLs, the nature of the security still has a long way to go. This paper proposes developing malicious URLs detection system which is used to identify new variants of known malware as well as to examine the presence of dangerous URLs seen in websites. The proposed study applies a heuristic based approach and collect URL features from the websites. Hence, the focus of this paper is to detect malicious website based on URL, the main contributions of this paper are the following:

a) The evaluation study applied URLs features for malicious and benign sample from Kaggle dataset.
b) The proposed PSO has improved the optimization of URLs features using tenfold cross-validation.
c) The proposed naïve bayes and SVM has increased the accuracy in classifying the optimized URLs features from malicious and benign websites applications.

The rest of the paper is organized as follows. In Section 2 discuses related works of the research. Section 3 describes the methodology which includes features optimization and general architecture. Section 4 evaluates the effectiveness of malicious URLs detection system. Lastly, Section 5 conclusion of this paper.

2. RELATED WORK

Malicious URLs contains vulnerabilities and poses a significant threat to the computer. This malicious website threat has become an important rising issue [17, 18, 19]. Many studies have been proposed for analyzing and detecting malicious URLs. Three types of approaches are used to detect malicious URLs which is static analysis, dynamic analysis, and heuristic analysis.

The static analysis determines the URLs whether malicious or benign based on the extracted source code. Mostly, the URLs that contain suspicious code will be assigned as a malicious website. In [20] examined the malicious websites based on HTML codes. They analyze the characteristic of malicious URLs to detect malicious or not. Their results show that their approach is resilient to code obfuscation and able to determine correctly whether the URLs is malicious or not. In [21] focused on drive-by download attack to detect malicious URLs by using traffic in a real network. They propose two-stage drive-by download attack detection mechanism which examines malicious URLs based on domain reputation and applying sandboxing approach to monitor the network based on URL and reduce the detection time. Based on the experiment, they achieved 94% of accuracy and able to reduce time more than 12 times compare in real computing traffic.

Many types of research concerned about the risk and impact on their computer when surfing website. In [18] proposed risk assessment to monitor the risk on URL by using the destination information when generating a short URL. By monitoring URL, any risky URL or risk over the threshold will be blocked to prevent malicious attack [19], especially from drive-by download through the short URL. In [20, 22] applied machine learning for detecting malicious websites. In [22] implemented three supervised machine learning techniques such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Naïve Bayes (NB). Beside the supervised machine learning, they also apply unsupervised machine learning technique to detect malicious websites such as Affinity Propagation and K-Means. Based on the experiment, their proposed produced 98% of accuracy for supervised machine learning and 96% of accuracy for unsupervised machine learning technique.

3. RESEARCH METHOD

This section describes the architecture of the experiment as it is important for executing the experiments. In the process for detecting malicious URLs, this paper applied optimization and machine learning approaches to optimize and train the sample. Optimizing and training sample are important in order to learn the behavior of malware and benign application.

Figure 1 presents the main components of the malicious URLs detection system. There are three phases in detection architecture including data collection, machine learning, and database. The data collection begins with crawling all the URL including malware and benign website applications. Then the data are processed through a features selection engine to collect relevant features for training purposes.

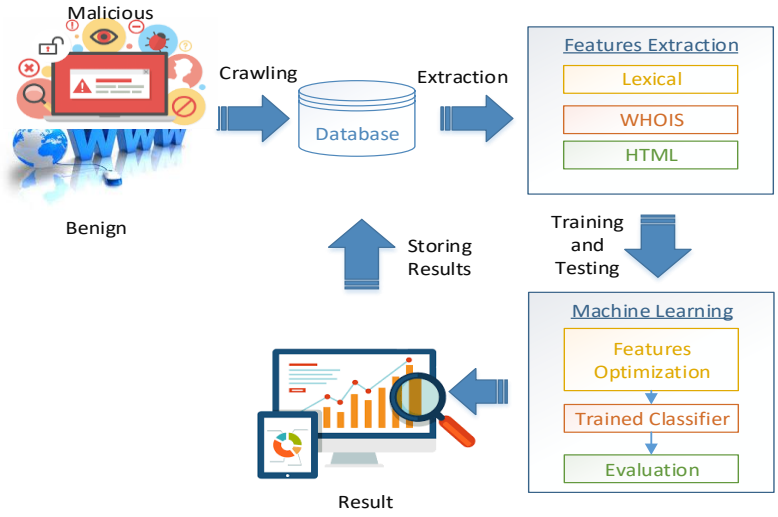

Figure 1. Malicious URLs detection architecture

3.1. Features Extraction and Selection

A relevant of features is significant to get the high performance for the machine learning [21, 23]. The criteria of features are important as to present the essential characteristics of the malicious website [24]. The process includes removing the noise and irrelevant features in the dataset. Table 1 shows the list of URLs features used in the experiment.

Table 1 lists the URL features for malicious URLs detection system. These features are important for the construction of the classification model, malicious URL detection process and identification of attack types. Here, malicious URLs and benign URLs features are classed in a binary number (0 or 1). It indicates 1 if the features are an exist in URLs and 0 if the features are a non-exist. Then this features are used to train, test and features optimization on WEKA.

Table 1. List of URLs Features

Features	Description
Token Count	The total number count of words in the URLs
Rank Host	The popularity ranking of the hostnames
Rank Country	The popularity ranking of the URLs (websites) among countries
ASNno	Autonomous System Number as the classifier for the IP of each URLs
Sec_sen_word_cnt	The security sensitive word count from the URLs
Avg_token_lenght	The total average number length count of the URLs
No_of_dots	The number of dots in the URLs
Length_of_url	The length of the URLs
Avg_path_token	The average number of the path for URLs

3.2. Machine Learning

This section aims to apply a machine learning approach for selecting the relevant features [25] used for detecting malicious URLs. In order to select the relevant features, an optimization approach is implemented to optimize the URLs features. This optimization approach could reduce the time for training,

testing and simplifying the malicious URLs detection system [20, 26]. Besides, it is also important for data processing [22]. Without a good knowledge of classification, it is difficult for malware analysis to identify relevant features for malicious URLs. Therefore, features optimization is the best approach to use for increase effectiveness of malicious URLs detection and improve accuracy. Hence, this study applied particle swarm optimization (PSO) for features optimization based on tenfold cross-validation.

Then, the performance optimization is compared with different classifiers in order to evaluate the effectiveness in malicious URLs detection system. Five machine learning classifiers, namely AdaBoost, Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Naïve Bayes and Random Forest are used for building the machine learning model in WEKA.

4. RESULTS AND ANALYSIS

In To evaluate the performance of the machine learning approach in detecting malicious URLs, the benign and malicious URLs application were mixed together for training and testing purposes. The training and testing models for machine learning, the parameter including the cross-validation needs to be set. Table 2 illustrates the detection performance of machine learning as seen in various categories of classifiers.

Table 2. Detection Performance of Machine Learning

Classifier	Accuracy	TPR	FPR	Precision	Recall	F-measure
Random Forest	97%	0.960	0.020	0.980	0.960	0.970
Naïve Bayes (This study)	99%	0.980	0.000	1.000	0.980	0.990
k-NN	97%	0.980	0.040	0.961	0.980	0.970
SVM(This study)	99%	0.980	0.000	1.000	0.980	0.990
AdaBoost	97%	0.960	0.020	0.980	0.980	0.970

Table 2 shows the detection performance of five classifiers for malicious URLs detection. The performance of each classifier is evaluated by six performance metrics such as accuracy, true positive rate (TPR), false positive rate (FPR), precision, recall, and f-measure. Table 2 indicates that Naïve Bayes, k-NN and SVM recorded highest TP Rate with 0.98 compared to another two algorithms which are Random Forest and AdaBoost, both recorded 0.96. This meant that the three algorithms have high sensitivity towards malicious data. Furthermore, both Naïve Bayes and SVM did not trace any false positives from the dataset since both recorded zeros for FPR. Other than that, both Naïve Bayes and SVM recorded the highest precision (1) which giving precise in predicting the malicious dataset. Hence, through those recorded results, the naïve Bayes and SVM present a better performance with 99% accuracy compared to the other classifiers. It is worth noting that machine learning with features optimization plays important role in identifying the relevant features in detecting malicious URLs.

5. CONCLUSION

This paper has presented the performance of the proposed approach in detecting malicious URLs. The proposed approach that implements the optimization has optimized the selection of URL features and the machine learning classifier has correctly classified the relevant malicious features. In the experiments, this paper considers applied real URL malware and benign samples application dataset. The experiment results show that the proposed approach recorded high accuracy in classifying the URLs malware samples.

ACKNOWLEDGEMENTS

This work was supported by Universiti Malaysia Pahang, under the Grant Faculty of Computer Systems and Software Engineering (FSK1000), RDU1803163.

REFERENCES

[1] S. G. Selvaganapathy, M. Nivaashini, and H. P. Natarajan, "Deep belief network based detection and categorization of malicious URLs," *Inf. Secur. J.*, vol. 27, no. 3, pp. 145–161, 2018.
[2] A. Firdaus, N. B. Anuar, M. F. A. Razak, and A. K. Sangaiah, "Bio-inspired computational paradigm for feature investigation and malware detection: interactive analytics," *Multimed. Tools Appl.*, 2017.
[3] D. R. Patil and J. B. Patil, "Feature-based Malicious URL and Attack Type Detection Using Multi-class Classification," *Int. J. Inf. Secur.*, vol. 10, no. 2, pp. 141–162, 2018.

[4] A. Kulkarni and L. L., "Phishing Websites Detection using Machine Learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 7, 2019.
[5] N. Jayakanthan and A. V. Ramani, "Classification Model to Detect Malicious URL via Behaviour Analysis," *Int. J. Comput. Appl. Technol. Res.*, vol. 6, no. 3, pp. 133–140, 2017.
[6] B. Li, G. Yuan, L. Shen, R. Zhang, and Y. Yao, "Incorporating URL embedding into ensemble clustering to detect web anomalies," *Futur. Gener. Comput. Syst.*, vol. 96, pp. 176–184, 2019.
[7] Muhammad Taseer Suleman and Shahid Mahmood Awan, "Optimization of URL-Based Phishing Websites Detection through Genetic Algorithms," *Autom. Control Comput. Sci.*, vol. 53, no. 4, pp. 333–341, 2019.
[8] N. M. M. Noor, S. Mohamad, Y. M. Saman, and M. S. Hitam, "Probabilistic knowledge base system for forensic evidence analysis," *J. Theor. Appl. Inf. Technol.*, vol. 59, no. 3, pp. 708–717, 2014.
[9] A. Firdaus, N. B. Anuar, M. F. A. Razak, I. A. T. Hashem, S. Bachok, and A. K. Sangaiah, "Root Exploit Detection and Features Optimization: Mobile Device and Blockchain Based Medical Data Management," *J. Med. Syst.*, vol. 42, no. 6, 2018.
[10] R. Zur, "12 Alarming Cyber Security Facts and Stats," *Cybrint*, 2018. [Online]. Available: https://www.cybintsolutions.com/cyber-security-facts-stats/. [Accessed: 03-Dec-2018].
[11] V. Chebyshev, F. Sinitsyn, D. Parinov, A. Liskin, and O. Kupreev, "IT threat evolution Q1 2018. Statistics," *Kaspersky Lab*, 2018. [Online]. Available: https://securelist.com/it-threat-evolution-q1-2018-statistics/85541/. [Accessed: 03-Dec-2018].
[12] A. Talalaev, "February 2018 Website Hacking Statistics," *WebARX Security*, 2018. [Online]. Available: https://www.webarxsecurity.com/website-hacking-statistics-2018-february/. [Accessed: 03-Dec-2018].
[13] K. Townsend, "18.5 Million Websites Infected With Malware at Any Time," *Security Week*, 2018. [Online]. Available: https://www.securityweek.com/185-million-websites-infected-malware-any-time. [Accessed: 03-Dec-2018].
[14] A. Firdaus, M. Faizal, A. Razak, and A. Feizollah, *The rise of " blockchain ": bibliometric analysis of blockchain study*, no. 0123456789. Springer International Publishing, 2019.
[15] F. Douksieh and L. Wenjuan, "Malicious Url Detection Using Convolutional Neural Network," *Int. J. Comput. Sci. Eng. Inf. Technol.*, vol. 7, no. 6, pp. 29–36, 2017.
[16] R. Wang, Y. Zhu, J. Tan, and B. Zhou, "Detection of malicious web pages based on hybrid analysis," *J. Inf. Secur. Appl.*, vol. 35, pp. 68–74, 2017.
[17] M. F. A. Razak, N. B. Anuar, F. Othman, A. Firdaus, F. Afifi, and R. Salleh, "Bio-inspired for Features Optimization and Malware Detection," *Arab. J. Sci. Eng.*, 2018.
[18] H. J. Mun and Y. Li, "Secure Short URL Generation Method that Recognizes Risk of Target URL," *Wirel. Pers. Commun.*, vol. 93, no. 1, pp. 269–283, 2017.
[19] W. I. S. W. Din, S. Yahya, R. Jailani, M. N. Taib, A. I. M. Yassin, and R. Razali, "Fuzzy logic for cluster head selection in wireless sensor network," *AIP Conf. Proc.*, vol. 1774, 2016.
[20] Y. T. Hou, Y. Chang, T. Chen, C. S. Laih, and C. M. Chen, "Malicious web content detection by machine learning," *Expert Syst. Appl.*, vol. 37, no. 1, pp. 55–60, 2010.
[21] C. M. Chen, J. J. Huang, and Y. H. Ou, "Efficient suspicious URL filtering based on reputation," *J. Inf. Secur. Appl.*, vol. 20, pp. 26–36, 2015.
[22] H. B. Kazemian and S. Ahmed, "Comparisons of machine learning techniques for detecting malicious webpages," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1166–1177, 2015.
[23] M. Akiyama, T. Yagi, T. Yada, T. Mori, and Y. Kadobayashi, "Analyzing the ecosystem of malicious URL redirection through longitudinal observation from honeypots," *Comput. Secur.*, vol. 69, pp. 155–173, 2017.
[24] S. B. Rathod and T. M. Pattewar, "A comparative performance evaluation of content based spam and malicious URL detection in E-mail," *2015 IEEE Int. Conf. Comput. Graph. Vis. Inf. Secur. CGVIS 2015*, pp. 49–54, 2016.
[25] M. Hazim, N. B. Anuar, M. F. Ab Razak, and N. A. Abdullah, "Detecting opinion spams through supervised boosting approach," *PLoS One*, vol. 13, no. 6, pp. 1–23, 2018.
[26] M. F. A. Razak, N. B. Anuar, R. Salleh, A. Firdaus, M. Faiz, and H. S. Alamri, "'Less Give More': Evaluate and zoning Android applications," *Meas. J. Int. Meas. Confed.*, vol. 133, pp. 396–411, 2019.